Performance Analysis of Mobility Prediction Based Proactive Wireless Caching

Yu Ye*, Ming Xiao*, Zhengquan Zhang†, and Zheng Ma†
*School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden
†Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Chengdu, China
Email: {yu9, mingx}@kth.se, zhang.zhengquan@hotmail.com, zma@swjtu.edu.cn

***Abstract*—We study a mobility prediction based proactive wireless caching scheme for two-tier cellular networks consisting of a base station (BS) tier and a device-to-device (D2D) tier. Two scenarios are considered: popular contents cached only at BSs, and popular contents cached at both BSs and MTs. We model user mobility as a Markov renewal process to predict user moving paths and residence time. Then we analyse the hit-rate performance to evaluate the presented schemes. By formulating content placement to maximize the hit-rate as optimization problems, we provide the optimal solution for the first scenario and develop a greedy mobility prediction based proactive wireless caching (MPPC) scheme for the second. Through analysis we show that the hit-rate achieved by MPPC is at least $\frac{\exp(1)-1}{\exp(1)}$ of the optimal hit-rate. The numeric results show that the MPPC can dramatically improve the hit-rate performance, compared with random caching and most popular caching (MPC) schemes. We show that the hit-rate achieved by MPPC outperformances MPC by 26% at most when MTs are not able to cache. Besides we present the impact of the moving speed on the hit-rate performance of MPPC for MTs.**

***Index Terms*—Proactive caching; mobility prediction; device-to-device communications**

I. Introduction

As one promising technology for the fifth-generation (5G) wireless networks and beyond, *proactive caching* can alleviate the heavy burdens on backhual links and reduce service delay, by proactively storing popular contents at base stations (BSs) and mobile terminals (MTs) [1]–[3]. User mobility is an important factor affecting popular content caching, which has attracted research attentions recently. In [4], a general framework on mobility-aware caching in content-centric wireless networks, was presented. In [5], [6], a storage allocation scheme for wireless caching in a two-tier heterogeneous network (HetNet) was studied, to minimize the probability of using macro BSs for content delivery, by taking user mobility into account. In [7], groups of mobile devices collaborate to exchange contents via BS assisted D2D communications, was studied. Deterministic and random caching strategies at MTs are analysed, and it is shown that the latter may be more realistic in networks with MT mobility. The effect of user mobility on the coverage probability of D2D networks with distributed caching, was studied in [8]. Finally in [9], the inter-contact time between MTs was considered, and a mobility-aware caching strategy was studied to maximize the percentage of requested data that can be delivered via D2D links.

Since MTs may move around, the design of proactive caching schemes for BSs and MTs should take both residence time and moving path into account, which however have not been fully explored in aforementioned works when studying the impact of user mobility on the performance of proactive caching. Therefore, we consider these two factors during user mobility, and study proactive wireless caching schemes for a two-tier cellular network. First, we model user mobility as a Markov renewal process to predict moving patterns, which consists of possible MT moving paths and corresponding residence time. Then for the scenario of caching at BSs, we present the optimal proactive caching scheme, by formulating the maximum hit-rate problem. Finally, we approximate user mobility through a temporal-spatial vector, and developed a heuristic algorithm mobility prediction based proactive caching (MPPC) to solve the maximum hit-rate problem for the scenario of caching at both BSs and MTs. The algorithm complexity and performance of MPPC are analyzed.

The rest of this paper is organized as follows. We present the system model of mobility prediction based proactive wireless caching in Section II. The optimal scheme and performance analysis under caching at BSs and caching at both BSs and MTs are studied in Sections III and IV, respectively. Numerical results are given to evaluate the performance of proposed proactive wireless caching schemes in Section V, followed by conclusions in Section VI.

II. System Model

Fig. 1 illustrates the system model of mobility prediction based proactive wireless caching, which consists of one central controller (CC), one BS tier and one D2D tier. The BS set is represented by $\mathcal{M} = \{1, 2, ..., M\}$, while $\mathcal{N} = \{1, 2, ..., N\}$ is for the MT set, and the content set is $\mathcal{F} = \{1, 2, ..., F\}$ respectively, where $M, N, F \in \mathbb{Z}^+$. BS m is equipped with caching unit which can store $\vartheta_m (\in \mathbb{Z})$ popular contents from $\mathcal{F}$, while MT n has a storage capacity $\vartheta_n^\dagger (\in \mathbb{Z})$. Without loss of generality, we assume that contents cannot be partitioned and must be stored as a whole at each BS and MT. Furthermore, we assume that the CC knows content preferences for all MTs.

Time is divided into multiple slots and each time slot is represented by T. The content preference for MT $n \in \mathcal{N}$ is $\boldsymbol{p}_n \in \mathbb{R}^{F\times 1}$. During time duration T, MT n requests contents drawn from $\boldsymbol{p}_n$ with a request arrival rate, which follows a Poisson distribution with mean λ_n. During time slot T, MTs may move from one BS to another. This leads to the uncertainty on where user request contents. To improve the hit-rate performance of requesting contents, the CC needs to

This work was supported by the EU Marie Sklodowska-Curie individual Fellowship under Grant 796426.

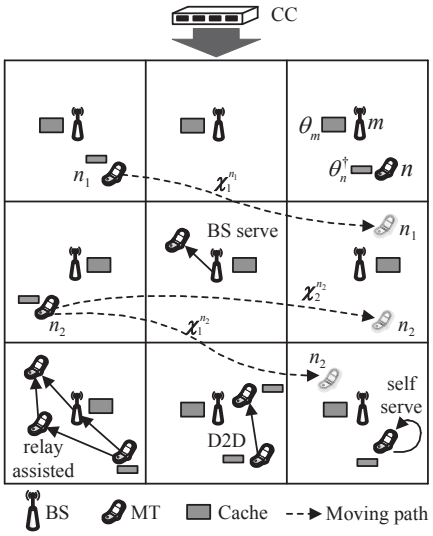

Fig. 1: System model of proactive wireless caching based on mobility prediction.

predict MT mobility to determine optimal content placements. Denote $\boldsymbol{S} \in \{0,1\}^{M\times F}$ as the content placement at BSs, where the element $S_{m,f} = 1$ means that the content f is stored at BS m during T, while $S_{m,f} = 0$ is for the case that content f is not available at BS m. In the same way, content placement at MTs is represented by $\boldsymbol{S}^\dagger \in \{0,1\}^{N\times F}$. The mobility prediction and proactive caching are accomplished at the beginning of T. We consider that the MT is associated to the nearest BS. Besides, the contents can be shared among MTs through D2D communications or relay assisted communications. We model MT n mobility as *Markov renewal process* [10], $\{(\mathcal{X}_i^n, \mathcal{T}_i^n) : i \geq 0\}$, to predict the moving path and the residence time at each BS, where $\mathcal{T}_i^n$ is the time instant of the i-th transition ($\mathcal{T}_0^n = 0$) and $\mathcal{X}_i^n \in \mathcal{M}$ is the state at the i-th transition. The transition probability of the embedded Markov chain for MT n is $\boldsymbol{P}^n \in \mathbb{R}^{M\times M}$. Let H_i^n denote the distribution of time that the semi-Markov process of MT n spends in state i before making a transition. The initial state for MT n is supposed to be $\mathcal{X}_0^n \in \mathcal{M}$, and the time that MT n has stayed at $\mathcal{X}_0^n$ before T is t_0^n.

III. Mobility Prediction Based Proactive Wireless Caching at BSs

In this section, we study proactive wireless caching scheme for the scenario that only BSs can store contents. For MT n, since we cannot know exactly how long n will spend at state $m \in \mathcal{M}$ within T, an efficient way to predict the residence time is to consider the average time evaluated as

$$\mu_m^n = \int_0^\infty x H_m^n(x)dx, \tag{1}$$

where $H_m^n(x)$ is the probability density function (pdf) of the residence time for MT n at state m. It means that MT n is expected to make a transition, when the time duration staying at state m exceeds μ_m^n.

We define the moving path set for MT n within time T as $\boldsymbol{X}^n = \{\boldsymbol{\mathcal{X}}_1^n, \boldsymbol{\mathcal{X}}_2^n, ..., \boldsymbol{\mathcal{X}}_{U^n}^n\}(U^n \geq 1)$ with initial state $\mathcal{X}_{u,0}^n = \mathcal{X}_0^n$ for any u, where the u-th ($1 \leq u \leq U^n$) path is $\boldsymbol{\mathcal{X}}_u^n = [\mathcal{X}_{u,0}^n, \mathcal{X}_{u,1}^n, ..., \mathcal{X}_{u,d_u^n}^n](d_u^n \geq 0)$. Note that MT n can be served by the same BS more than one times in T, and hence $\boldsymbol{\mathcal{X}}_u^n$ may be a multiset.

Since MT n stays at $\mathcal{X}_0^n$ for t_0^n time, the 1-st transition for all paths in $\boldsymbol{X}^n$ is forecasted to occur at time instant

$$\mathcal{T}_{u,1}^n = \int_{t_0^n}^\infty x H_{\mathcal{X}_{u,0}^n}^n(x)dx. \tag{2}$$

$\mathcal{T}_{u,1}^n$ is the same among all possible paths and $\mathcal{T}_{u,0}^n = 0$. Considering the u-th path, the instant time of the i-th transition is predicted to be

$$\mathcal{T}_{u,i}^n = \mathcal{T}_{u,i-1}^n + \mu_{\mathcal{X}_{u,i-1}^n}^n, \quad 2 \leq i \leq d_u^n. \tag{3}$$

Note that the last transition will take place before the end of T. This indicates $\mathcal{T}_{u,d_u^n-1}^n \leq T$. $\boldsymbol{\mathcal{R}}_u^n = \{\mathcal{R}_{\mathcal{X}_{u,0}^n}^n, \mathcal{R}_{\mathcal{X}_{u,1}^n}^n, ..., \mathcal{R}_{\mathcal{X}_{u,d_u^n}^n}^n\}$ denotes the residence time for the u-th path $\boldsymbol{\mathcal{X}}_u^n$. Since $\mathcal{R}_{\mathcal{X}_{u,i}^n}^n = \mathcal{T}_{u,i}^n - \mathcal{T}_{u,i-1}^n (1 \leq i \leq d_u^n - 1)$, and considering the content delivery time τ, the residence time is derived as

$$\mathcal{R}_{\mathcal{X}_{u,i}^n}^n = \begin{cases} \left[\mathcal{T}_{u,1}^n - \tau\right]_0^+, & i = 0 \\ \left[\mu_{\mathcal{X}_{u,i}^n}^n - \tau\right]_0^+, & 1 \leq i \leq d_u^n - 1 \\ \left[T - \sum_{j=0}^{d_u^n-1} \mathcal{R}_{\mathcal{X}_{u,j}^n}^n - \tau\right]_0^+, & i = d_u^n \end{cases}, \tag{4}$$

where $y = [a]_0^+$ confines the value $y = a$ when $a \geq 0$, and $y = 0$ while $a < 0$. If $\mathcal{T}_{u,1}^n \geq T$, there is no transition for MT n during T then $U = 1$ and $d_u^n = 0$. The probability of the moving path $\boldsymbol{\mathcal{X}}_u^n$ for MT n within T equals to

$$P\left(\boldsymbol{\mathcal{X}}_u^n\right) = \prod_{i=0}^{d_u^n-1} P_{\mathcal{X}_{u,i}^n, \mathcal{X}_{u,i+1}^n}. \tag{5}$$

Combining (4) and (5), we can predict the average residence time, $\bar{\mathcal{R}}_i^n$, at state m during T, as

$$\bar{\mathcal{R}}_m^n = \sum_{u=1}^{U} P(\boldsymbol{\mathcal{X}}_u^n) \sum_{j=0}^{d_u^n} \mathbb{1}(\mathcal{X}_{u,j}^n = m) \mathcal{R}_{\mathcal{X}_{u,j}^n}^n, \quad m \in \mathcal{M}, \tag{6}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Now we take the request density and content preference of MT n into account. The average hit-rate $\bar{r}_{m,f}^n$ that MT n requests content f at BS m is expressed as

$$\bar{r}_{n,f}^m = \lambda_n \bar{\mathcal{R}}_m^n p_{n,f} S_{m,f}, \quad f \in \mathcal{F}. \tag{7}$$

Therefore, the optimal proactive caching placement at BSs can be formulated as the following optimization problem

$$\begin{aligned} \max_{\boldsymbol{S}} \quad & r_{\text{hit-rate}} = \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \sum_{f\in\mathcal{F}} \bar{r}_{n,f}^m \\ s.t. \quad & \sum_{f\in\mathcal{F}} S_{m,f} \leq \vartheta_m, \quad m \in \mathcal{M}. \end{aligned} \tag{P1}$$

Then, by defining weight vector as $\boldsymbol{w}_m \triangleq [w_{m,1}, ..., w_{m,F}]$ where $w_{m,f} = \sum_{n\in\mathcal{N}} \lambda_n \bar{\mathcal{R}}_m^n p_{n,f}$, we give the following theorem to obtain the optimal solution for Problem P1.

Theorem 1. *The optimal $r_{\text{hit-rete}}$ for Problem P1 is achieved when BS $m \in \mathcal{M}$ stores contents greedily according to the descending order of $\boldsymbol{w}_m$ until all the caching space is occupied.*

Proof. Problem P1 is a linear programming problem and can be divided into M sub-problems. The m-th sub-problem is to optimize variables $S_{m,f}(f \in \mathcal{F})$ with weight $w_{m,f}$. It is optimal to store the content which has larger weight. Thus optimal $r_{\text{hit-rate}}$ can be achieved by storing contents according to the descending order of $\boldsymbol{w}_m$ at $m \in \mathcal{M}$. □

IV. Mobility Prediction Based Proactive Wireless Caching at BSs and MTs

In this section, we will study the proactive wireless caching scheme when both BSs and MTs can store contents. To utilize the storage of MTs more efficiently and improve the hit-rate of the whole system, the proactive caching strategy both for BSs and MTs should be optimized.

A. Problem Formulation and Heuristic Solution

To investigate the moving pattern of MTs, we divide T into $K(\in \mathbb{Z}^+)$ equal time slices $\frac{T}{K}$. Without loss of generality, we choose K to satisfy $K > \max\{d_u^n | 1 \le u \le U^n, n \in \mathcal{N}\}$. For the u-th $(1 \le u \le U^n)$ moving path of MT n, we introduce a vector $\boldsymbol{\mathcal{V}}_u^n$ to involve both *temporal* and *spatial* information as

$$\boldsymbol{\mathcal{V}}_u^n = \left[\mathcal{V}_{u,1}^n, \mathcal{V}_{u,2}^n, ..., \mathcal{V}_{u,K}^n\right], \tag{8}$$

where $\mathcal{V}_{u,k}^n \in \mathcal{M}(1 \le k \le K)$. If MT n stays at state m at the end of the k-th time slice, then we denote $\mathcal{V}_{u,k}^n = m$. It is clear that $\boldsymbol{\mathcal{V}}_u^n$ is an expansion of $\boldsymbol{\mathcal{X}}_u^n$ on the time horizon. $\boldsymbol{\mathcal{V}}_u^n$ and $\boldsymbol{\mathcal{X}}_u^n$ consist of same elements with different dimensions. To determine $\boldsymbol{\mathcal{V}}_u^n$, we should derive the discrete representation $\boldsymbol{\mathcal{K}}_u^n = [\mathcal{K}_{u,0}^n, ..., \mathcal{K}_{u,d_u^n+1}^n]$ of transition time for path u. We set $\mathcal{K}_{u,0}^n = 0$ and $\mathcal{K}_{u,d_u^n+1}^n = K$,

$$\mathcal{K}_{u,i}^n \triangleq \left\lfloor \frac{K}{T}\mathcal{T}_{u,i}^n + \frac{1}{2} \right\rfloor, \quad 0 \le i \le d_u^n. \tag{9}$$

Taking the content delivery time τ into account, we define the feasible discrete transition time as

$$\widetilde{\mathcal{K}}_{u,i}^n \triangleq \left\lfloor \frac{K}{T}\left[\mathcal{T}_{u,i}^n - \min\{\tau, \mathcal{T}_{u,i}^n - \mathcal{T}_{u,i-1}^n\}\right]_0^+ + \frac{1}{2} \right\rfloor, \quad 1 \le i \le d_u^n. \tag{10}$$

Then with (9) and (10), the element of $\boldsymbol{\mathcal{V}}_u^n$ can be evaluated as

$$\mathcal{V}_{u,k}^n = \begin{cases} \mathcal{X}_{u,i-1}^n, & \mathcal{K}_{u,i-1}^n + 1 \le k \le \widetilde{\mathcal{K}}_{u,i}^n \\ 0, & \widetilde{\mathcal{K}}_{u,i}^n + 1 \le k \le \mathcal{K}_{u,i}^n \end{cases}, \quad 1 \le i \le d_u^n + 1. \tag{11}$$

The derivation of $\boldsymbol{\mathcal{V}}_u^n$ is shown in Fig. 2. Thus we have obtained $\boldsymbol{\mathcal{V}}_u^n$, which is an approximated description of the moving path u for MT n. Since T is divided into K slices, $\mathcal{K}_{u,i}^n$ and $\widetilde{\mathcal{K}}_{u,i}^n$ are not the actual transition time. The accuracy of $\boldsymbol{\mathcal{V}}_u^n$ representing path u can be enhanced by enlarging K. However this will increase the dimension of $\boldsymbol{\mathcal{V}}_u^n$, and leads to high computation complexity.

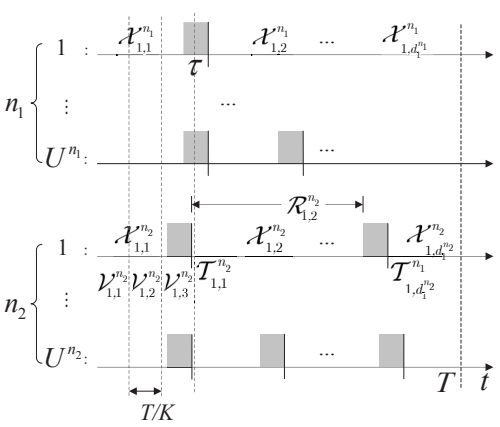

Fig. 2: The hit-rate comparison between proposed solutions and MPC and RC schemes where $\gamma = 0.9$.

Considering the contents cached at BSs and local storage of MT n, and defining $S_{0,f} \triangleq 0(f \in \mathcal{F})$, the hit-rate of MT n requesting content f in the k-th time slice can be evaluated as

$$\bar{r}_{n,f}^k = \lambda_n \frac{T}{K} p_{n,f} \sum_{u^1=1}^{U^1} ... \sum_{u^N=1}^{U^N} \prod_{i \in \mathcal{N}} P(u^i) \max\left\{ \underbrace{S_{n,f}^\dagger}_{\text{local}}, \underbrace{S_{\mathcal{V}_{u^n,k}^n, f}}_{\text{BS}}, \underbrace{\mathbb{1}\{\mathcal{V}_{u^j,k}^j = \mathcal{V}_{u^n,k}^n\} S_{j,f}^\dagger}_{\text{MT}} | j \in \mathcal{N}\backslash n \right\}, \tag{12}$$

where the set $\mathcal{N}\backslash n$ includes the elements in $\mathcal{N}$ except MT n.

The optimal proactive wireless caching is to optimize content placement to achieve the maximum total hit-rate. Thus, we formulate the maximum total hit-rate problem as

$$\begin{aligned} \max_{\boldsymbol{S}} \quad & r_{\text{hit-rate}} = \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{k=1}^{K} \bar{r}_{n,f}^k \\ s.t. \quad & \sum_{f \in \mathcal{F}} S_{m,f} \le \vartheta_m, \quad m \in \mathcal{M}, \\ & \sum_{f \in \mathcal{F}} S_{n,f}^\dagger \le \vartheta_n^\dagger, \quad n \in \mathcal{N}. \end{aligned} \tag{P2}$$

We will solve P2 in a heuristic way. To combine the proactive caching for BSs and MTs, we combine $\mathcal{N}$ and $\mathcal{M}$ into a node set $\hat{\mathcal{N}} = \{1, ..., N, N+1, ..., N+M\}$. This is because BSs can be seen as a special type of MTs which neither request nor move in T. Hence for the nodes $N+1 \le n \le N+M$ which represent BSs, we denote $\lambda_n = 0$, $\boldsymbol{p}_n = \boldsymbol{0}$, $U^n = 1$, $\boldsymbol{\mathcal{V}}_u^n = (n-N) \cdot \boldsymbol{1}_{1 \times K}$, $P(U^n) = 1$, storage capacity as $\vartheta_n^\dagger = \vartheta_{n-N}$, and content placement as $\boldsymbol{S}_n^\dagger$. We first calculate the contact time $\mathcal{C}_{n,i}$ of MT n and node $i \in \hat{\mathcal{N}}\backslash n$, which is the time duration that they spend at the same states

$$\mathcal{C}_{n,i} = \sum_{u^n=1}^{U^n} \sum_{u^i=1}^{U^i} P(u^n) P(u^i) \frac{T}{K} \left(\boldsymbol{\mathcal{V}}_{u^n}^n \oplus \boldsymbol{\mathcal{V}}_{u^i}^i\right) \left(\boldsymbol{\mathcal{V}}_{u^n}^n \oplus \boldsymbol{\mathcal{V}}_{u^i}^i\right)^T, \tag{13}$$

where A^T is the transpose of A, and $\boldsymbol{x} \oplus \boldsymbol{y}$ is the exclusive-or (EOR) operation of vector $\boldsymbol{x}$ and $\boldsymbol{y}$. The elements of $\boldsymbol{x}$ and $\boldsymbol{y}$

need not necessarily to be 0 or 1, and $(\boldsymbol{x}\oplus\boldsymbol{y})(\boldsymbol{x}\oplus\boldsymbol{y})^T$ counts the same elements from $\boldsymbol{x},\boldsymbol{y}$. By denoting $C_{n,n}=T(n\in\hat{\mathcal{N}})$, we define the utility function of caching content f at node $n\in\hat{\mathcal{N}}$ with respect to all the other MTs as

$$\mathcal{U}(n,f)=\left(1-S^{\dagger}_{n,f}\right)\sum_{i\in\mathcal{N}}\lambda_i p_{i,f}\mathcal{C}_{n,i}\left(1-S^{\dagger}_{i,f}\right). \quad (14)$$

Based on this utility function, we develop a greedy algorithm shown in Algorithm 1 to determine the proactive caching strategy $\boldsymbol{S}^{\dagger}\in\{0,1\}^{(N+M)\times F}$.

Algorithm 1: Mobility prediction based proactive wireless caching at BSs and MTs (MPPC)

Input: $\{\lambda_n,\boldsymbol{p}_n,\vartheta^{\dagger}_n|n\in\hat{\mathcal{N}}\},\{\mathcal{C}_{i,j}|i\neq j,i,j\in\hat{\mathcal{N}}\}$.
1: Initialization: $\boldsymbol{S}^{\dagger}\leftarrow\mathbf{0}_{(N+M)\times F}$, $\hat{\boldsymbol{\vartheta}}^{\dagger}\leftarrow\mathbf{0}_{1\times(N+M)}$, calculate $\{\mathcal{U}(n,f)|n\in\hat{\mathcal{N}},f\in\mathcal{F}\}$ according to (14).
2: **while** $\boldsymbol{\vartheta}^{\dagger}-\hat{\boldsymbol{\vartheta}}^{\dagger}\neq\mathbf{0}$ **do**
3: $(n^*,f^*)=\arg\max_{n,f}\mathcal{U}(n,f)$;
4: **if** $\mathcal{U}(n^*,f^*)=0$ **then**
5: break;
6: **else if** $\hat{\vartheta}^{\dagger}_{n^*}<\vartheta^{\dagger}_{n^*}$ **then**
7: $S^{\dagger}_{n^*,f^*}\leftarrow 1$; $\hat{\vartheta}^{\dagger}_{n^*}\leftarrow\hat{\vartheta}^{\dagger}_{n^*}+1$;
8: Update $\{\mathcal{U}\{n,f^*\}|n\in\hat{\mathcal{N}}\}$;
9: **else if** $\hat{\vartheta}^{\dagger}_{n^*}=\vartheta^{\dagger}_{n^*}$ **then**
10: $\mathcal{U}(n^*,:)\leftarrow\mathbf{0}_{1\times F}$;
11: **end if**
12: **end while**
Output: $\boldsymbol{S}^{\dagger}$.

B. Analysis of MPPC

In MPPC, we choose to store content f^* at node n^* at each iteration, which has the largest utility, if node n^* still has available storage capacity. The complexity of the initialization step of Algorithm 1 is assumed to be $\mathcal{O}(1)$. Considering the worst case for the sorting and updating operation, the complexity of MPPC can be evaluated as $\mathcal{O}(1+\sum_{n\in\hat{\mathcal{N}}}\vartheta^{\dagger}_n[(N+M)F\log(N+M)F+(2+N+M)])\approx\mathcal{O}(\sum_{n\in\hat{\mathcal{N}}}\vartheta^{\dagger}_n(N+M)F\log(N+M)F)$.

For problem P2, the greedy algorithm MPPC provides an effective solution.

Theorem 2. *When $K\to\infty$, denote the optimal hit-rate of P2 as r^*, and the hit-rate obtained by MPPC as r, then we have*

$$\frac{r}{r^*}\geq\frac{\exp(1)-1}{\exp(1)}. \quad (15)$$

Proof. The hit-rate $r_{\text{hit-rate}}$ in P2 is a function of content placement strategy $\boldsymbol{E}\triangleq\{e_{n,f}|S_{n,f}=1,n\in\hat{\mathcal{N}},f\in\mathcal{F}\}$, which is denoted as $r_{\text{hit-rate}}=r(\boldsymbol{E})$. Now consider any two caching stratgies $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ satisfying $\boldsymbol{E}_1\subseteq\boldsymbol{E}_2\subseteq\boldsymbol{E}_f$, where $\boldsymbol{E}_f\triangleq\{e_{n,f}|n\in\hat{\mathcal{N}},f\in\mathcal{F}\}$. Adding the content $j\in\boldsymbol{E}_f-\boldsymbol{E}_2$ to both $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$, the total hit-rate have

$$r(\boldsymbol{E}_1\cup\{j\})-r(\boldsymbol{E}_1)\geq r(\boldsymbol{E}_2\cup\{j\})-r(\boldsymbol{E}_2). \quad (16)$$

This is because adding the content j may serve more MT requests when fewer contents have been stored. Thus the hit-rate r is a submodular set function. When $K\to\infty$, the contact time of (13) is accurate. Besides in each iteration of MPPC, the content which can increase the hit-rate most is stored. Hence from the results in [11], we can conclude that the greedy MPPC can achieve an approximation as

$$\frac{r}{r^*}\geq\frac{\exp(1)-1}{\exp(1)}. \quad (17)$$

□

Besides, the proposed MPPC can also be applied to solving problem P1.

Corollary 1. *Setting $\vartheta^{\dagger}_n=0(1\leq n\leq N)$ and $K\to\infty$ in MPPC, an optimal hit-rate, same as that achieved in Theorem 1, will be obtained.*

Proof. While $\vartheta^{\dagger}_n=0(1\leq n\leq N)$, the outcome of Algorithm 1 is the content placement for BSs $\mathcal{M}$. Besides when $K\to\infty$, the contact time of (13) for $n\in\mathcal{N}$ and $i\in\mathcal{M}$ will be the same as that of (4). From Theorem 1, the optimal content placement for BS $m\in\mathcal{M}$ is to store contents greedily according to weights $\boldsymbol{w}_m$. Hence the greedy algorithm MPPC will achieve the optimal hit-rate. □

V. Numerical Results

In this section, we will illustrate the numerical results of the presented schemes and give discussions in detail. Most popular caching (MPC) and random caching (RC) [4] are chosen as benchmarks. For MPC, the CC does not consider the movement of MTs. The weight for content f seen by BS m is $\hat{w}_{m,f}=\sum_{n\in\mathcal{N}}\mathbb{1}\{\mathcal{X}^n_0=m\}\lambda_n T p_{n,f},\ f\in\mathcal{F}$. Then, the BS m caches contents according to the descending order of $\hat{\boldsymbol{w}}$ until all caching capacity is occupied. While MT n caches contents greedily following the descending order of $\boldsymbol{p}_n$. For RC, contents cached at BSs are chosen randomly, as well as MTs.

In our simulation setup, the number of BSs is fixed to $M=4$, the contents is fixed to $F=100$, and the amount of MTs is $N=10$. We normalize the length of each content as $l=1$. The storage capacity for different BSs are the same as ϑ. Besides different MT has unique storage capacity $\vartheta^{\dagger}$. We generate the vector $\boldsymbol{p}_n$ to follow a Zip-f like distribution [12], where the shape parameter γ among MTs is the same. The time duration is set to $T=10$ min and $K=100$. While the transmission time of requested contents is $\tau=0.1$ min. The distribution of residence time for each MT follows uniform distribution. The request density λ_n, initial state $\mathcal{X}^n_0$ and time t^n_0 are generated randomly for MT n, as well as the embedded Markov chain $\boldsymbol{P}^n$.

A. Hit-rate Comparison

Fig. 3 shows the hit-rate for the presented schemes, MPC, and RC. The results show that the hit-rate increases with enlarging BS storage capacity ϑ. Meanwhile, for fixed ϑ, the hit-rate also increases with $\vartheta^{\dagger}$ due to content sharing through D2D communications. The hit-rate achieved by caching at both BSs and MTs is better than that of MPC and RC. This is because user mobility is taken into account when the CC determines the proactive caching scheme for BSs and MTs.

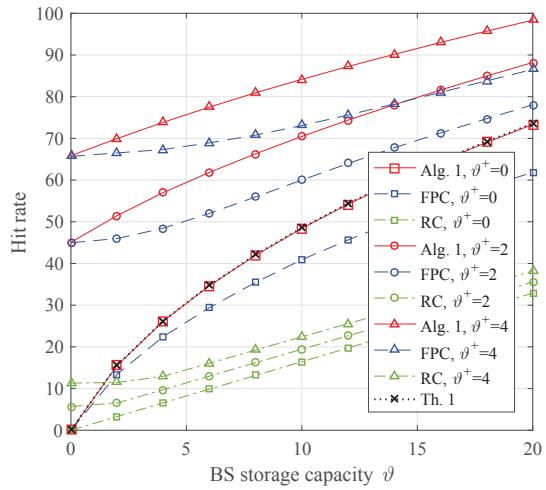

Fig. 3: The hit-rate comparison between proposed solutions and MPC and RC schemes where $\gamma = 0.9$.

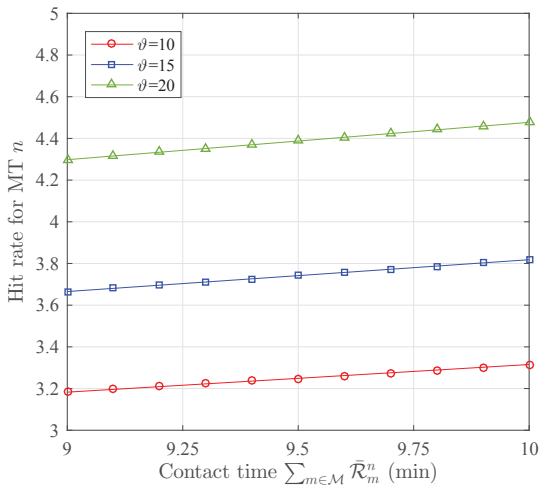

Fig. 5: Hit rate for MT n with different contact time, where $\gamma = 0.9$ and $\vartheta^\dagger = 0$.

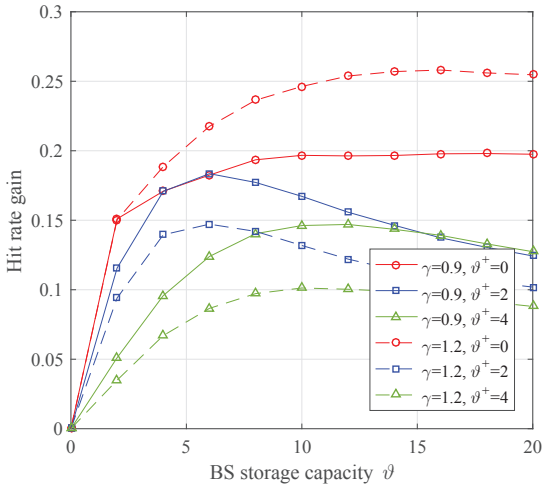

Fig. 4: Hit rate gain of MPPC compared with MPC scheme with different shape parameter γ.

Besides when $\vartheta^\dagger = 0$, the hit-rate obtained by proposed MPPC is the same as that of Th. 1, as shown by Corollary 1.

In Fig. 4, we compare the hit-rate gain for the scheme that caches contents at both BSs and MTs, and MPC, with different shape parameter γ. The hit-rate gain is defined as $\mathcal{G}_{\text{hit-rate}}^{\text{MPPC, MPC}} = \frac{r_{\text{hit-rate}}^{\text{MPPC}} - r_{\text{hit-rate}}^{\text{MPC}}}{r_{\text{hit-rate}}^{\text{MPPC}}}$. We can observe that when the shape parameter $\gamma = 0.9$ and $\vartheta = 0$, the hit-rate gain will increase sharply at the low ϑ region and then keeps at the same level at the medium and high region. However a larger capacity $\vartheta^\dagger$ does not necessarily guarantee a better hit-rate gain. We show that the hit-rate gain of $\vartheta^\dagger = 2$ is higher than that of $\vartheta^\dagger = 4$. This demonstrates that the benefit of proactive caching involving MT mobility will decay with growing $\vartheta^\dagger$. When the shape parameter gets to $\gamma = 1.2$, the popular contents become more deterministic. While MT caching is enable, the hit-rate gain is larger than that of $\gamma = 0.9$, since more requests can be satisfied when the storage capacity of the BS keeps the same. But with growing $\vartheta^\dagger$, the hit-rate gain shows a similar trends with that of $\gamma = 0.9$. The difference is that for fixed $\vartheta^\dagger > 0$, the hit-rate gain of $\gamma = 1.2$ is smaller than that of $\gamma = 0.9$.

B. Impact of the Mobility Speed

To invesitgate the impact of mobility speed on the hit-rate, we define the speed of MT based on its moving pattern. From (4) it can be concluded that the faster MT $n(\in \mathcal{N})$ moves, the shorter contact time $\sum_{m\in\mathcal{M}} \bar{\mathcal{R}}_m^n$ it has with BSs in set $\mathcal{M}$. Here we do not consider the impact of relative moving speed between MTs. Fig. 5 shows the hit-rate of MT n while MTs are not able to cache. As the speed of n increases, which subsequently decreases the contact time with BSs, the hit-rate performance of MT n degrades. The first reason is that as the contact time of MT n decreases, the content preference of MT n has less impact on the caching decision of BSs. Hence the probability of popular contents for MT n available at BSs is reduced. Besides from (7), the times for MT n receiving requested contents successfully from BSs reduces when the moving speed of MT n increases.

VI. Conclusions

We studied the proactive wireless caching schemes for two-tier cellular networks, by taking user mobility into account. We modeled user mobility as a Markov renewal process to predict user moving path and residence time. Then, we considered caching only at BSs and caching at both BSs and MTs, and formulated the optimal content placements as maximizing hit-rate optimization problems. Furthermore, we developed the optimal solution and heuristc algorithm for these two problems respectively. Finally, numerical results show that the proposed mobility prediction based proactive wireless caching scheme (MPPC) can significantly improve the hit-rate. Besides we show that the hit-rate performance for an MT achieved by MPPC degrades with increasing moving speed.

REFERENCES

[1] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131-139, February 2014.

[2] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82-89, Aug. 2014.

[3] D. Malak, M.A-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[4] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: modeling and methodology," *IEEE Commun. Mag.*, vol. 54, pp. 77-83, August 2016.

[5] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Istambul, Turkey, Jul. 2013, pp. 1017-1021.

[6] K. Poularakis and L. Tassiulas, "Code, cache and deliver on the move: a novel caching paradigm in hyper-dense small-cell networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 675–687, Mar. 2017.

[7] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665-3676, July 2014.

[8] S. Krishnan, and H. S. Dhillon, "Effect of user mobility on the performance of device-to-device networks with distributed caching," *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 194–197, Apr. 2017.

[9] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief "Mobility-Aware Caching in D2D Networks," *arXiv preprint arXiv:1606.05282*, 2017.

[10] J.-K. Lee and J. C. Hou, "Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application," in *Proc. 7th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, 2006, pp. 85-96.

[11] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, "An analysis of approximations for maximizing submodular set functions-I," *Math. Program.*, vol. 14, no. 1, pp. 265-294, Dec. 1978.

[12] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," *IEEE INFOCOM*, New York, 1999, pp. 126-134.